

# A regression model for estimation of dew point pressure from down-hole fluid analyzer data

Ahmed Alzahabi<sup>3</sup> · Ahmed El-Banbi<sup>2</sup> · A. Alexandre Trindade<sup>1</sup> · Mohamed Soliman<sup>3</sup>

Received: 30 March 2016 / Accepted: 27 November 2016 / Published online: 4 January 2017  
© The Author(s) 2017. This article is published with open access at Springerlink.com

**Abstract** Accurate knowledge of dew point pressure is important in understanding and managing gas condensate reservoirs. Without a correct assessment of dew point pressure, an accurate description of phase changes and phase behavior cannot be achieved. Numerous models for predicting gas condensate dew point pressure from surface fluid data have been proposed in the literature. Some of these require knowledge of the full composition of the reservoir fluid (based on laboratory experiments), while others only need field parameters that are relatively easy to obtain. This paper presents a new model for predicting the dew point pressure from down-hole fluid analyzer data. Such data are now measured (usually in real time) while obtaining down-hole fluid samples. The new model predictions give a quick estimation of dew point pressure for wet gas and gas condensate reservoirs. Since it relies only on down-hole measured data, the model provides an estimate of dew point pressure without the need for laboratory

analyses. During down-hole fluid sampling, the model can be used to ensure whether the sample is still in single phase, or whether the dew point was crossed during the sampling operation. An early estimate of dew point pressure is also valuable in designing further tests for gas condensate wells. The new model, constructed using a fluid database of nearly 700 gas condensate samples, was devised using sophisticated statistical/machine learning methods, and attained a mean absolute relative error value of 2% for predicting the logarithm of pressure. In comparison with other dew point estimation models (that use surface fluid data), the chosen model was found to attain a similar level of accuracy when tested on samples not used in the model building phase.

**Keywords** Regression model · Dew point pressure · Down-hole fluid analyzer data · Gas condensate reservoirs

## Abbreviations

API	API gravity of stock-tank condensate
CFA	Compositional Fluid Analyzer
DFA	Down-hole fluid analyzer
GCR	Gas condensate ratio, SCF/STB
IFA	In situ Fluid Analyzer
LFA	Live Fluid Analyzer
Mwt <sub>C7+</sub>	Molecular weight of heptane plus
OFA	Optical fluid analyzer
Pd	Dew point pressure, psia
Ppr	Pseudo-reduced pressure
Psp	Primary separator pressure, psia
Rsp	Primary separator gas–oil ratio, SCF/STB
T	Reservoir temperature, F
Tpr	Pseudo-reduced temperature
Tsp	Primary separator temperature, F

✉ Ahmed Alzahabi  
amalzahabi@uh.edu

Ahmed El-Banbi  
abanbi@gmail.com

A. Alexandre Trindade  
alex.trindade@ttu.edu

Mohamed Soliman  
msoliman@Central.UH.EDU

<sup>1</sup> Department of Mathematics and Statistics, Texas Tech University, Broadway and Boston, Lubbock, TX 79409-1042, USA

<sup>2</sup> Cairo University, Giza, Egypt

<sup>3</sup> Department of Petroleum Engineering, University of Houston, UH Energy Research Park (ERP), Houston, TX 77023, USA

$Y_i$	Component mole fraction
$Y_{C7+}$	Mole Percentage of heptane plus
$\rho_{C7+}$	C7 plus density, g/cc
$\gamma_{gsp}$	Separator gas specific gravity
$\gamma_g$	Gas specific gravity (air = 1)
$\gamma_{gR}$	Reservoir gas specific gravity
$\gamma_{C7+}$	Specific gravity of heptane plus
$\gamma_{cond}$	Condensate gas specific gravity
MARE	Mean absolute relative error
AIC	Akaike information criterion
BIC	Bayesian information criterion
MSE	Mean squared error
$R^2$	A criterion for correlation coefficient of a linear regression
Beta	Model coefficients
$n$	No of samples
$\varepsilon$	Residual error term

## Introduction

Dew point pressure is needed to characterize wet gas and gas condensate reservoir fluids. Industry practitioners often rely on measuring the dew point (among other phase behavior properties) using laboratory experiments. In the absence of laboratory data, dew point estimation models (correlations) are usually available to estimate the dew point with varying accuracy (discussed below). These models are based on either knowledge of fluid composition or knowledge of some surface fluid properties data (e.g., GCR, API of stock tank oil, and reservoir temperature). For many years, the oil and gas industry has been actively developing several tools to measure real-time in situ fluid composition and properties. To the best of our knowledge, we are unaware of any simple dew point estimation model based on the down-hole fluid composition measurements except the one we present here. Such model, when available, will allow rapid evaluation of dew point pressure before measuring it in the laboratory with common techniques and will have several applications.

Numerous models for predicting gas condensate dew point pressure have been derived from large databases in the literature. There are essentially two types of published models for estimating dew point pressure in gas condensate reservoirs. One type of models uses detailed compositional analysis that requires laboratory PVT experiments, while another type uses easily measured parameters from production tests and down-hole temperature as inputs. Marruffo et al. (2002), Nemeth and Kennedy (1966), Elsharkawy (2001, 2011), Shokir (2008), Olds et al. (1944), and Godwin (2012) require detailed compositional

analysis. Marruffo et al. (2002) used nonlinear regression to fit appropriate models and build their model, applying statistical tools such as residual analyses and cross-plots. The developed model required information from production tests, but did not require the knowledge of gas condensate composition. The original total PVT data sample size they used was 148. After the process of selection and validation, the database was reduced to 114 data points. Nemeth and Kennedy (1966) used 579 data points from 480 different hydrocarbon systems to develop a model that predicts dew point pressure with an average deviation of 7.4%. The model input parameters are variables measured in the laboratory with composition in mole fraction for CH<sub>4</sub> through C<sub>7</sub>H<sub>16</sub>, N<sub>2</sub>, CO<sub>2</sub>, H<sub>2</sub>S, and molecular weight and specific gravity of heptane plus fraction. Their work is regularly quoted in many of the more recent models. Elsharkawy (2001) developed a physically sound empirical method for predicting dew point pressure based on routinely measured gas analysis and reservoir temperature. In total, 340 measurements of dew point pressure were used, resulting in a model with an absolute average of 7.68%. The model included the effect of all variables such as temperature, gas condensate composition, and properties of the plus fraction on dew point pressure. Elsharkawy (2001, 2011) discussed the two types of dew points applicable to any hydrocarbon mixtures. The first dew point occurs when dry gas is compressed to the point that liquid starts to form. The second called retrograde or condensate dew point and occurs when a gas mixture containing heavy hydrocarbons in its solution is depressurized until liquid forms. He presented models for dew point prediction based on gas composition and reservoir temperature. His database included 340 data points. Shokir (2008) used genetic programming to develop a model for dew point pressure prediction from 245 gas condensate systems. The developed model uses the full composition of the gas (CH<sub>4</sub> through C<sub>7</sub>H<sub>16</sub>+, N<sub>2</sub>, CO<sub>2</sub>, and H<sub>2</sub>S mole fractions) in addition to the molecular weight of the heptanes plus fraction, and reservoir temperature. He also tested his model against other published models. Olds et al. (1944) studied the behavior of six symmetrically chosen mixtures from Paloma field for a range of parameters (temperatures ranged from 100 to 250 °F and pressures up to 5000 psia). They studied the influence of pressure and temperature on the composition and the retrograde gas dew point, and implemented a graphical examination in order to obtain a chart correlating volumetric and phase behavior with the composition of the system and temperature. Godwin (2012) used data from the literature and developed a dew point estimation model based on gas composition, reservoir temperature, and properties of the heptanes plus fraction. A total of 259 out of 273 data points were selected

to build the new model, and 14 data points were used for testing.

On the other hand, a literature review for the models that take as input field data that are easily measured finds the following: Humoud and Al-Marhoun (2001), Ovalle et al. (2005) and Al-Dhamen and Al-Marhoun (2011). Humoud and Al-Marhoun (2001) developed another model based on available field data from 74 PVT reports. They correlated the dew point pressure of a gas condensate fluid directly with its reservoir temperature, pseudo-reduced pressure and temperature, primary separator gas–oil ratio, the primary separator pressure and temperature, and relative densities of separator gas and heptanes plus specific gravity. The average error for this model was 4.33%. Ovalle et al. (2005) used readily available field data to calculate the dew point pressure. Their database contained 615 points. Their model is based on initial producing gas condensate ratio from the first-stage separator, initial API of the stock tank liquid, specific gravity of the initial reservoir gas, and reservoir temperature. Nonparametric approaches for estimating optimal transformations of data were used to obtain the maximum correlation between observed variables. Al-Dhamen and Al-Marhoun (2011) developed a new model to predict dew point pressure for gas condensate reservoirs, using nonparametric approaches and artificial neural networks. Their results were based on a total number of 113 data samples obtained from constant mass expansion experiments from fields in the Middle East.

### Down-hole fluid analysis

The process of obtaining real-time analysis of down-hole characteristics passed through many stages of development, starting with IFA and ending with DFA. Fingerprinting in fluid characterization became an important topic receiving wide attention with direct application on improving the quality of fluid samples. Many tools (e.g., OFA, LFA, CFA, and IFA) capable of detecting in situ variation of different fluids were developed over the years (1991/2001/2003/2007), (Mullins et al. 2009; Elshahawi et al. 2007; Xian et al. 2006). The development of these tools was to address several production problems (e.g., sizing of facilities, well placement, completions equipment, and production prediction). According to Betancourt et al. (2004, 2007), the Composition Fluid Analyzer (CFA) is a tool which has a sensor for performing fluorescence spectroscopy by measuring light emission in the green and red ranges of the spectrum after excitation with blue light. It was originally introduced in order to track phase transitions in gas condensate sampling. In situ Fluid Analyzer (IFA), based on optical absorption methods, can provide the mass fractions of three hydrocarbon molecular groups:  $\text{CH}_4$ ,  $\text{C}_2\text{H}_6\text{--C}_5\text{H}_{12}$  and  $\text{C}_6\text{H}_{14+}$ , as well as  $\text{CO}_2$ , in real

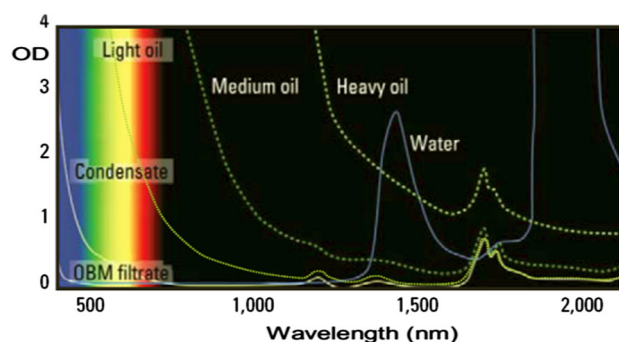
time at down-hole conditions. It can also track the gas condensate (by simply dropping temporarily the sampling pressure below the saturation pressure of the fluid so as to observe the change in the fluorescence signal that will occur with dew formation at the dew point pressure).

With the development of down-hole optical fluid analyzer (DFA), more capabilities were added to down-hole fluid analysis. DFA has become an increasingly utilized technology in wireline logging as it enables fluid characterization by creating a down-hole fluid log (versus depth along the hydrocarbon column). In multi-well applications, DFA can help in addressing fluid distribution and variation inside the reservoir, and in identification of reservoir compartments. The basic outputs from DFA measurements are weight percentages of  $\text{CO}_2$ ,  $\text{CH}_4$ ,  $\text{C}_2\text{H}_6$ ,  $\text{C}_3\text{H}_8$ ,  $\text{C}_5\text{H}_{12}$ , and  $\text{C}_6\text{H}_{14+}$ , in addition to live fluid density.

Mullins et al. (2009) showed that in the case of large fluid conditional variations and compartmentalization, DFA can be used as a tool to help in delineating these variations in a cost-effective manner. They introduced the example shown in Fig. 1 as an identifying fingerprint among different fluids. Analysis of the oil peak at a wavelength of 1700 nm gives the dissolved methane content. Therefore, it can be used for the tracking of density variations and discontinuities in fluid properties.

In the form of an optimized wireline logging tool, DFA is used in the Gulf of Mexico and different areas of the world for detecting hydrocarbon variations and reducing uncertainty in varied compositions cases. Compartmentalization can also be detected by these tools (Betancourt et al. 2007).

With the increased application of these down-hole fluid analysis tools, valuable compositional information (for hydrocarbon groups in weight percent) becomes available to reservoir and production engineers. In this paper, we present a new dew point estimation model that is different from the other models available in the literature, as it is based on down-hole fluid analyzer data. The correlation is thus capable of predicting the dew point pressure for a wide range of wet gases and gas condensate fluids without the



**Fig. 1** Visible near-infrared spectra of oilfield fluids, after Mullins et al. 2009

need for full laboratory compositional analysis, production data, or production test information.

## Methodology

### Fluids database

McCain (1994) characterized different fluid properties and introduced widely acceptable criteria to differentiate among the five reservoir fluid types. According to McCain's criteria, we collected fluid data (covering wide range of gas properties) from different reservoirs located in different regions of the world (with around 17% of the data from the Middle East). Part of the database came from the literature, especially the data presented by Nemeth (1966), which was extensively used in developing most available dew point pressure prediction models from surface data.

The database contained 667 complete (without missing values) laboratory gas condensate samples. We manipulated the data to be in the format of the output of down-hole fluid analyzer tools. We divided the data into two groups. The first group included the data where complete laboratory analyses were performed (Table 1) and consisted of 99 complete samples. The second group consisted of the remaining 568 samples which included compositional data and some basic information (Table 2). The full database

included gas condensate samples with reservoir gas gravity ranging from 0.558 to 1.86, primary separator gas gravity of 0.56–1.42, field stock tank liquid gravity of 37.0–67.7, condensate gas ratio of 0.63–232 STB/MMscf, separator temperature from 19.9 to 176 °F, separator pressure 33.20–2581.7 psia, C7+ specific gravity from 0.69 to 0.85, C7+ mole percent from 0 to 24.23, reservoir temperatures of 143.8–347 °F, and dew point pressure 1429–11,656 psia.

### Development of a new empirical model

In developing the model, we considered only the 667 complete gas condensate samples in our database (no missing variable values). In the model to predict dew point pressure (the output or dependent variable), the following pool of independent (or input) variables was considered: temperature, CO<sub>2</sub>, CH<sub>4</sub>, C<sub>2</sub>H<sub>6</sub>, C<sub>3</sub>H<sub>8</sub>, C<sub>4</sub>H<sub>10</sub>, C<sub>5</sub>H<sub>12</sub>, and C<sub>6</sub>H<sub>14</sub>+ mole%. The model building procedure entailed the following steps.

1. building a database for gas samples;
2. making quality checks on the data samples;
3. filtering the samples;
4. converting mole% to weight% for all samples based on molecular weight of each component to match the output of the down-hole fluid analyzer data;

**Table 1** Data ranges for the complete PVT experiments gas samples (99 samples)

Laboratory measured parameter	Minimum	25th percentile	50th percentile (Median)	75th percentile	Maximum
Reservoir gas gravity	0.558	0.765	0.89	1.08	1.86
Primary separator gas gravity	0.56	0.67	0.72	0.77	1.42
Field stock tank liquid gravity (°API)	37.00	43.49	48.55	54.50	67.70
GCR, STB/MMscf	0.63	42	120	207	232
Field gas–oil ratio, scf/STB	1147	4778	9207	19,944	159,8940
C7+ density g/cc	0.69	0.80	0.81	0.82	0.85
C7+ Mwt	101	144	158	180	353
Reservoir temperature, <i>T</i> (°F)	143.8	223	260	296	347
<i>T</i> <sub>sep</sub> , °F	19.9	81	99	119	176
<i>P</i> <sub>sep</sub> , psia	33	255	439	624	2581
Composition (mole%)					
Methane	9.19	66.24	73.60	82.81	99.84
Ethane	0.05	6.34	7.96	9.28	15.04
Propane	0.04	2.54	3.81	4.96	8.60
Butane	0.02	0.60	1.18	1.80	6.04
Pentane	0	0.65	1.14	2.94	3.92
Hexane	0	0.48	0.86	1.31	5.54
Heptane plus	0	2	5	8	24.23
Hydrogen sulfide	0	0	0	0	18.00
Carbon dioxide	0	0.64	2.6	5.32	67.08
Nitrogen	0	0.2	0.49	0.75	9.86

**Table 2** Data ranges for the composition available PVT gas samples (568 samples)

Laboratory measurement Parameters	Minimum	25th percentile	50th percentile (Median)	75th percentile	Maximum
Reservoir gas gravity	0.733	0.77	0.79	0.80	0.87
C7+ Mwt	106	128	147	160	235
Reservoir temperature, $T$ (°F)	40	168	212	246	320
Composition (mole%)					
Methane	3.49	74.59	82.55	88.91	96.68
Ethane	0.37	3.73	5.04	6.94	15.13
Propane	0.11	1.47	2.43	4.08	10.90
Butane	0.17	0.97	1.59	2.77	37.50
Pentane	0	0.45	0.86	1.52	07.10
Hexane	0	0.36	0.64	1.16	08.71
Heptane plus	0	1.53	2.76	5.23	13.56
Hydrogen sulfide	0	0.00	0.00	0.00	29.86
Carbon dioxide	0	0.28	0.62	1.37	91.92
Nitrogen	0	0.00	0.28	0.66	43.22

- PVT data compositions were lumped back to emulate the down-hole fluid analyzer output compositions; and
- checking interrelationships among the variables and removing poor predictor variables (e.g.,  $C_2H_6$ ).

Some exploratory analyses were carried out to find the best scale, transformations, and importance of the predictor variables (step 6 above). A moderate amount of correlation was found among most of the variables. These preliminary analyses led us to consider a multiple linear regression model with the following variables, where we have assigned them symbols in order to more easily discuss the results.

$y = \log$  (Dew Point Pressure)

$x_1 = \text{temperature}$

$x_2 = \log(CO_2 + 0.1)$

$x_3 = \log(CH_4)$

$x_4 = \log(CH_4)^2$

$x_5 = \log(C_3H_8 - C_5H_{12})$

$x_6 = \log(C_4H_{14}+)$

The units of  $CO_2$ ,  $CH_4$ ,  $C_2H_6$ ,  $C_3H_8$ ,  $C_5H_{12}$ , and  $C_6H_{14}+$  are in weight percent, temperature is in degrees Fahrenheit, and pressure is in psia.  $C_3H_8-C_5H_{12}$  denotes the weight percent of the group  $C_3H_8$  through  $C_5H_{12}$ . All logarithm ( $\log$ ) values denote *natural log* (base  $e$ ).

The scatter plots below the diagonal in the composite matrix plot in Fig. 2 give an idea of the pairwise relationships among the variables. The red line is a local linear smoother through the cloud of points. The blue values in the upper part of the matrix plot are the corresponding values of the correlation coefficients between each pair of variables, with font size proportional to the absolute value of the correlation. Thus, the largest correlation (0.98)

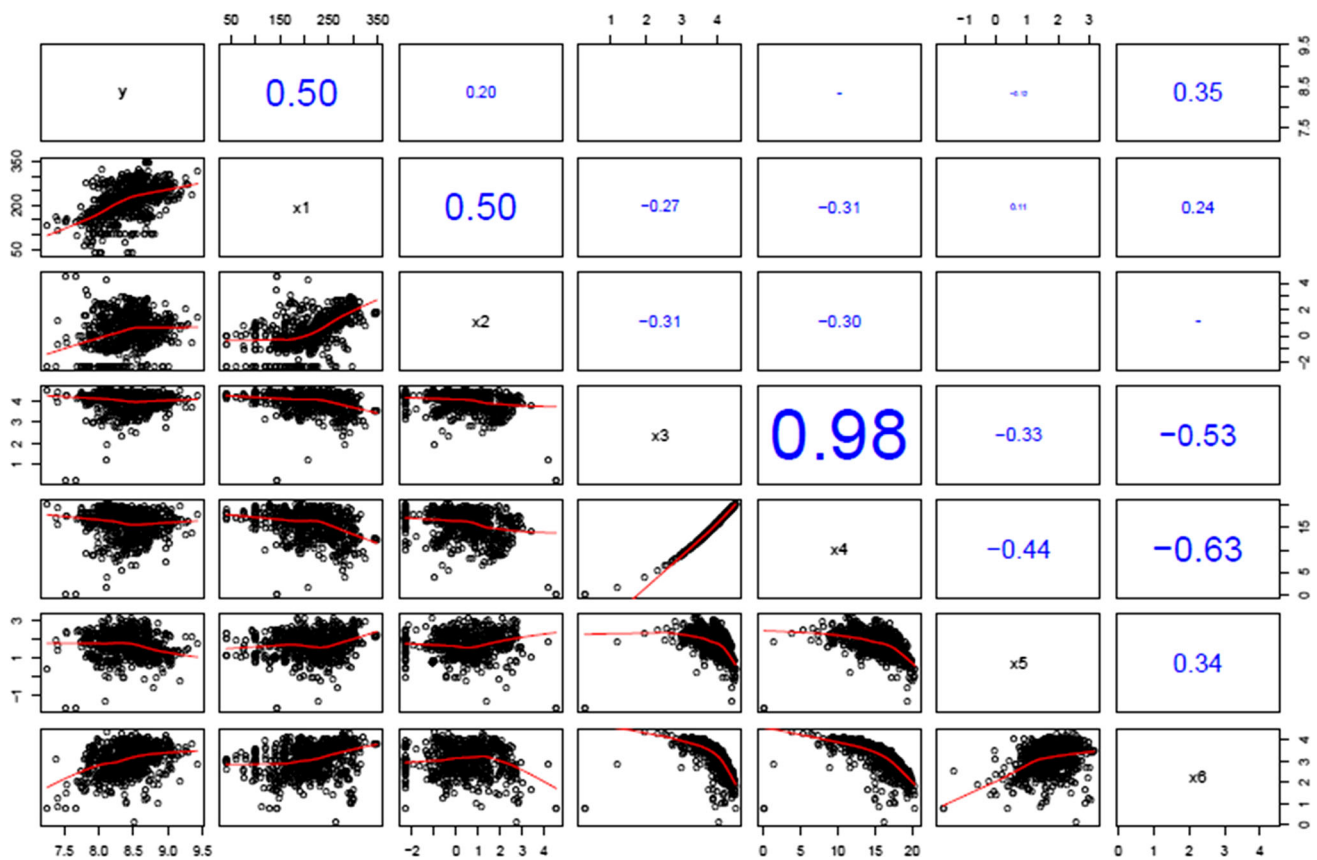
occurs between  $x_3$  and  $x_4$  (not surprising since  $x_4 = x_3^2$ ), and the smallest ( $-0.0018$  and too small to be visible) between  $x_2$  and  $x_5$ . Note that  $x_1$  (temperature) is the most important predictor of pressure since the two have a correlation coefficient of 0.50, while  $x_3$  is the least important (correlation between  $y$  and  $x_3$  is 0.004).

A multiple linear regression model was predicted via standard statistical methods to the database of  $n = 667$  well samples, resulting in the model listed in Table 3 and shown in Eqs. (3) and (4). Three criteria were used to select an appropriate model (AIC, AICc, BIC); however, the “best” model identified by each of these criteria is usually not the same (Burnham and Anderson 2002). The search involved considering models of the form:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_6 x_6 + \cdots + \beta_{12} x_{12} + \cdots + \beta_{56} x_{56} + \epsilon \quad (1)$$

where  $\epsilon$  is the usual residual noise term in a regression model. The  $\beta_i$  are model parameters to be estimated from our data. This was done by searching across all possible combinations of variables and their pairwise interactions. For example, the variable  $x_{34} = x_3 * x_4$  denotes the product of  $x_3$  and  $x_4$ , and is called the interaction between  $x_3$  and  $x_4$ . This resulted in a pool of 21 potential predictors: the 6 single variables  $\{x_1, \dots, x_6\}$ , plus a total of 15 interaction terms  $\{x_{12}, \dots, x_{56}\}$ . With all combinations of 21 variables, the number of possible candidate models that can be formed is  $2^{21} \approx 2.1$  million. (This can be understood by realizing that we have the option of whether or not to select each of the 21 predictors  $\{x_1, \dots, x_{56}\}$  for inclusion in the model.) The size of the model space to be searched over is thus extremely large. The complexity of this search was made possible by using sophisticated statistical software, namely the R





**Fig. 2** Scatter plot of variables used in the regression model

**Table 3** Coefficients of the chosen model

Coefficient	Estimate	SE	<i>P</i> value
$\beta_0$	19.1109840	1.9899250	0.000
$\beta_1$	-0.0679165	0.0104606	0.000
$\beta_2$	-0.0162705	0.1414375	0.908
$\beta_3$	-6.6190184	1.3607809	0.000
$\beta_4$	0.5104139	0.4082668	0.212
$\beta_5$	1.1398989	0.2306464	0.000
$\beta_6$	0.6263451	0.0487792	0.000
$\beta_{13}$	0.0371260	0.0055541	0.000
$\beta_{14}$	-0.0048367	0.0007388	0.000
$\beta_{23}$	0.0573708	0.0263501	0.030
$\beta_{26}$	-0.0565329	0.0160504	0.000
$\beta_{34}$	0.0794272	0.0471590	0.093
$\beta_{35}$	-0.1985207	0.0501634	0.000
$\beta_{56}$	-0.1334765	0.0241433	0.000

package glmulti (R Core Team 2016) which implements a genetic algorithm search over large model spaces.

Basic measures of model goodness of fit are  $R^2$ , mean squared error (MSE) which is the estimate of the noise variance  $\sigma^2$ , and mean absolute relative error (MARE),

defined as follows. If  $y_i$  and  $\hat{y}_i$  are, respectively, the observed and the model-predicted values for the  $i$ th value of  $y$ ,  $i = 1, \dots, n$ , then

$$\text{MARE} = \frac{1}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (2)$$

Note that  $n$  is the sample size and in our case  $n = 667$ . It is well known that  $R^2$  (proportional of variability in  $y$  explained by the model) will increase and both MSE and MARE will decrease, as more variables are included in the model, despite the importance of these predictor variables. Thus, an over-parameterized model (too many predictors) will have very high/low values of these measures, accordingly, and will seemingly do very well in sample, but will do poorly out of sample. The use of model selection tools based on information criteria such as AIC, AICc, and BIC tends to avoid the over-fitting problem (Burnham and Anderson 2002).

### Performance of the chosen model

The coefficient estimates for the best model according to the AICc criterion, are displayed in Table 3. For example,  $\beta_0 = 19.11$  and  $\beta_1 = -0.0679$ . The standard error (Std.

Error) column is an estimate of the variability of the estimate and can be used to assess the uncertainty associated via a formal hypothesis test. The  $P$  value in Table 3 is the result of testing if the corresponding parameter is equal to 0. For example, for  $\beta_0$  the  $P$  value of 0.000 means that the estimate of  $\beta_0$  is significantly different from 0, whereas the  $P$  value of 0.908 means that the estimate of  $\beta_2$  is not significantly different from 0. The other commonly used criteria (AIC, BIC) arrived at models that were very similar to this one.

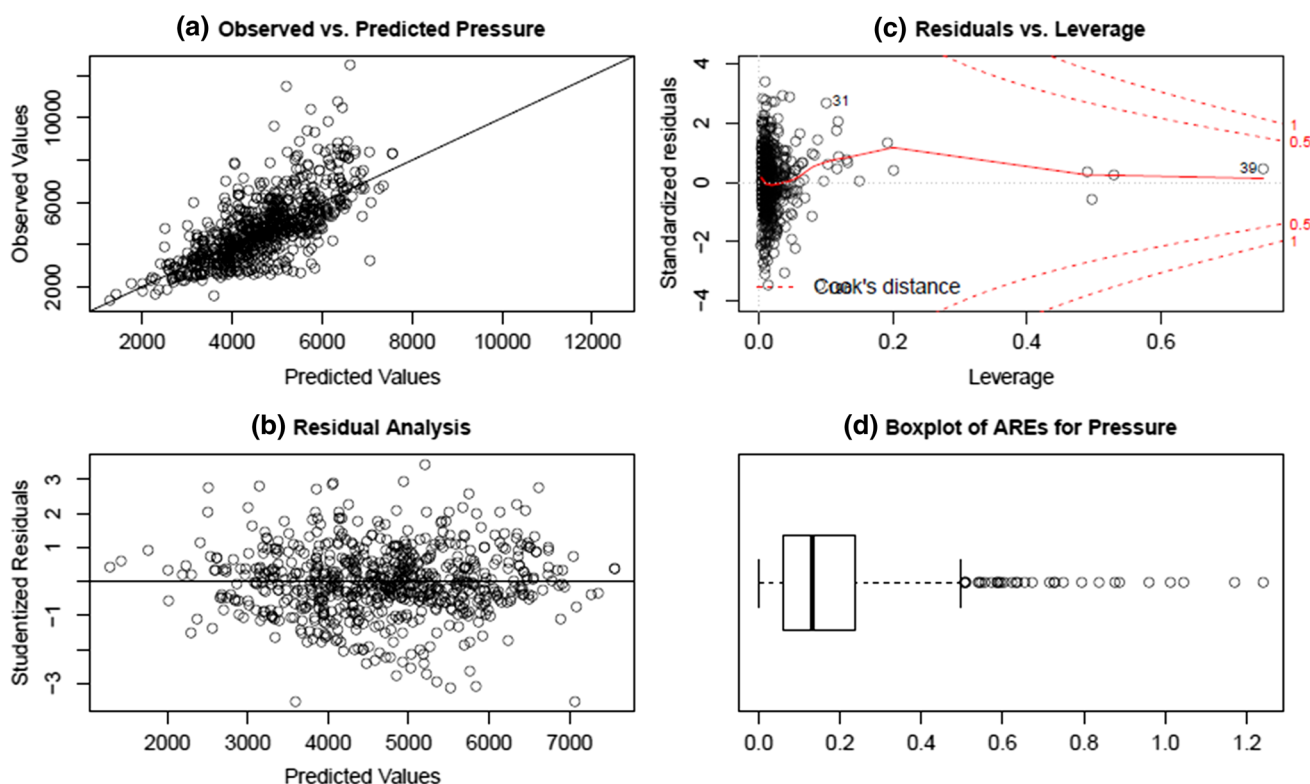
$$P_d = e^X \quad (3)$$

$$\begin{aligned} x = & 19.1109840 - 0.067916 \times \text{temperature} - 0.0162705 \\ & \times \log(\text{CO}_2 + 0.1) - 6.6190184 \\ & \times \log(\text{CH}_4) + 0.5104139 \times \log(\text{CH}_4)^2 + 1.1398989 \\ & \log(\text{C}_3\text{H}_8 - \text{C}_5\text{H}_{12}) \\ & + 0.6263451 \times \log(\text{C}_6\text{H}_{14}+) + 0.0371260 \times \text{temperature} \\ & \times \log(\text{CH}_4) - 0.0048367 \\ & \times \text{temperature} \times \log(\text{CH}_4)^2 + 0.0573708 \times \log(\text{CO}_2 + 0.1) \\ & \times \log(\text{CH}_4) \\ & - 0.0565329 \times \log(\text{CO}_2 + 0.1) \times \log(\text{C}_6\text{H}_{14}+) \\ & + 0.0794272 \times \log(\text{CH}_4) \\ & \times \log(\text{CH}_4)^2 - 0.1985207 \times \log(\text{CH}_4) \\ & \times \log(\text{C}_3\text{H}_8 - \text{C}_5\text{H}_{12}) - 0.1334765 \\ & \times \log(\text{C}_3\text{H}_8 - \text{C}_5\text{H}_{12}) \times \log(\text{C}_6\text{H}_{14}+) \end{aligned} \quad (4)$$

This model has  $R^2 = 0.54$ ,  $\text{MSE} = 0.23$ , and  $\text{MARE} = 0.0209$  (or approximately 2%) on the transformed log (pressure) scale, if measured on the original pressure scale; however, the MARE increases to about 17%. Standard diagnostic analysis shows that this model fits well, and the normality assumption on  $\varepsilon$  is reasonable (see Fig. 3). Figure 3a plots the model-predicted versus observed values of pressure and shows that there is generally close agreement. Figure 3b shows essentially the same information, but on a horizontal line, where the vertical axis is now the difference observed minus predicted (the residuals). Figure 3c shows a standard diagnostic used to identify data points that are not well fit by the model, and it does not indicate the presence of any overly problematic points in this case. Figure 3d shows a graphical summary (boxplot) of all 667 absolute relative errors (AREs). The vertical line inside the box of the boxplot, located at around 0.13, signifies that the median ARE is 13%. Overall, Fig. 3 suggests that the model provides a good fit to the data.

### Validation of the chosen model

The kind of modeling problem at hand is termed supervised learning in machine learning terminology, which has seen an explosion in activity in the last two decades. The most successful and theoretically sound approaches to solve this



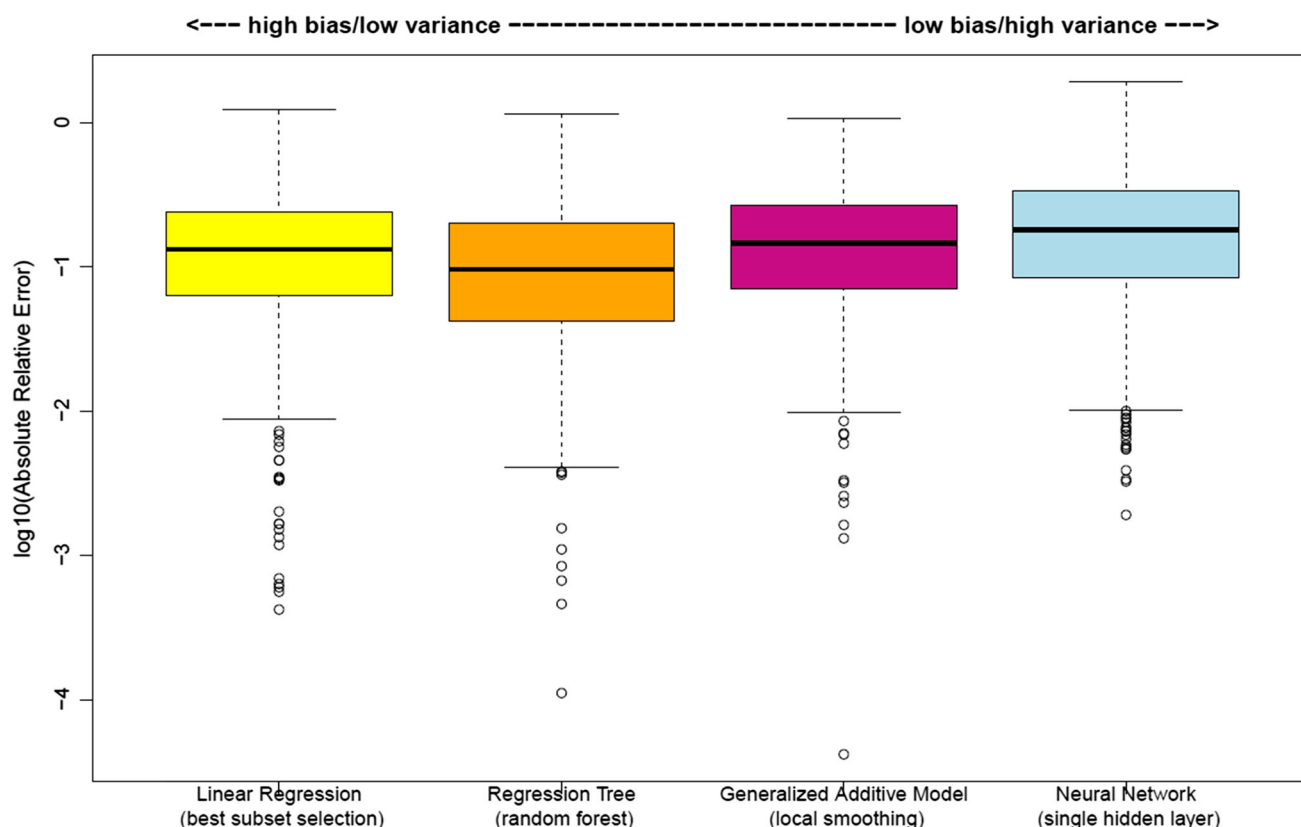
**Fig. 3** Diagnostic analysis of the chosen model

problem have recently been compiled by Hastie et al. (2009). They span the gamut of statistical methods from the high bias/low variance, e.g., linear regression, principal components (PCA), partial least squares (PLS), and least absolute shrinkage and selection operator (LASSO), to the low bias/high variance, e.g., splines, local smoothing, and neural networks. Roughly in the middle of this bias/variance trade-off dilemma, one finds regression tree-based models and extensions (bagging, boosting, random forests) to be some of the best predictive methods on a variety of different problems.

For the data set at hand, sparsity seeking and shrinkage inducing methods such as PCA, PLS, and LASSO are not really appropriate given the small number of predictors involved (only 6). Rather, more important is the capturing of complex nonlinear relationships with the output variable (dew point pressure) and interactions among the predictors. Thus, and in order to ensure we were considering all the best possible models, we decided to compare a variety of methods, restricting our attention to the following 4 classes: (1) linear regression with up to two-way interactions and all subsets search using a consistent information criterion such as BIC; (2) regression trees and the computationally intensive resampling-based extensions such as bagging, boosting, and random forests; (3) generalized additive models with individual predictor functions

estimated via splines and local smoothers; (4) feed-forward neural networks with a single hidden layer. Details of these methods can be found in Hastie et al. (2009). (Note that method 1 was the strategy used to arrive at the chosen model in Table 3).

In order to determine which of these methods should actually be employed here, we used the tried and tested paradigm of  $K$ -fold cross-validation, with the best general recommendation at present being something like  $K = 5$  or  $K = 10$  (Hastie et al. 2009). With fivefold cross-validation, we randomly split the data into 5 equal portions (fivefold), use fourfold to train the model, and use the remaining onefold to test. The absolute relative error (ARE) measure described above was used to evaluate the predictive ability of a given model. Thus, for any given training/test data combination, approximately 530 data points are used to fit the model and predict the remaining 130 points. Any decisions and selection of tuning parameters pertaining to a given candidate model must be made on a case-by-case basis for each of the fivefold, using the training/test set paradigm. The absolute difference between the observed and predicted values of pressure at these 130 points is then divided by the observed value, resulting in 130 ARE values. This exercise was repeated for each of the fivefold, so that each method yields 667 ARE values. Only the best performing model in each of the 4 classes described above



**Fig. 4** Boxplots of log base 10 ARE values for each of the 4 classes of models used to validate the chosen model



was considered. Figure 4 displays a statistical summary (boxplot) of log base 10 of the 667 ARE values pertaining to each of these 4 optimal models. (A boxplot extends approximately from the minimum value to the maximum value, with a box around the middle 50% of the data.)

We see that all methods perform similarly, with the random forest model (a type of regression tree) having a slight edge, and linear regression a close second. However, whereas it is straightforward to write down an equation for the linear regression model, this is infeasible for the regression tree-based random forest model, since it is a combination of thousands of trees, each tree being a sequence of yes/no questions about the predictors that must be answered sequentially in order to arrive at the appropriate predicted value. For this reason, and due to the fact that the difference in predictive ability between the two models is small, we have chosen to report only the linear regression model in this paper. However, our recommendation is that any future work should carefully consider regression trees.

## Results and discussion

As already mentioned, there are two types of published correlations for dew point pressure prediction. Some correlations use detailed compositional data while others use relatively easily measured parameters from production tests and fluid data as inputs. A review of the literature of dew point prediction models revealed the main 9 published correlations listed in Table 4.

Table 5 lists the main input parameters and number of PVT data points used for developing each of the published dew point pressure correlations. All the published models use fluid data and production parameters that are based on surface values, while the new model presented here uses down-hole data. The model was used to predict the 99 samples in Table 1 that were left out of the model building procedure. However, recall that our model uses down-hole data, whereas the other available models cannot use these types of input variables. This is an important fact for fast dew point pressure estimation in the field while sampling the fluid and before going to the laboratory.

**Table 4** Published dew point pressure correlations

Correlations requiring detailed compositional analysis	Correlations not requiring detailed composition
Marruffo et al. (2002)	Humoud and Al-Marhoun (2001)
Nemeth and Kennedy (1966)	Ovalle et al. (2005)
Elsharkawy (2001, 2011)	Al-Dhamen and Al-Marhoun (2011)
Shokir (2008)	
Olds et al. (1944)	
Godwin (2012)	

**Table 5** Published correlations and their required input parameters

Correlation	Number of points in data set	Input data
Olds et al. (1944)	N/A	$T, P$
Nemeth and Kennedy (1966)	579	$Y \text{ mol}\%, C_1-C_{7+}, M_{C_{7+}}, T, \rho_{7+}$
Elsharkawy (2001)	340	$Y \text{ mol}\% C_1-C_{7+}, M_{C_{7+}}, T, \rho_{7+}$
Humoud and Al-Marhoun (2001)	74	$R_{sp}, T, T_{sp}, P_{sp}, \gamma_{sp}, \gamma_{C_{7+}}$
Marruffo et al. (2002)	80	$YC_{7+}, T, API, CGR$
Ovalle et al. (2005)	615	$API, R_{sp}, \gamma_{gR}$
Shokir (2008)	245	$MC_7 +, T, Y \text{ mol}\% (C_1, CO_2, H_2S, N_2, C_{7+})$
Al-Dhamen and Al-Marhoun (2011)	113	$T, MC_7 +, \rho_{7+}, \gamma_{cond}, GOR, \gamma_{gs}$
Godwin (2012)	259	$T, Y \text{ mol}\% (C_1, C_2, \dots, C_6, C_{7+}), MC_7 +, \rho_{7+}$

The value of the new model proposed here lies in its simplicity and relative accuracy for the database used in this work. More importantly, it is based on down-hole fluid data that are becoming more available in today's applications of fluid sampling and fluid characterization. Unlike many of those available dew point models (correlations), the new model does not require information obtainable from the laboratory or production test data. Therefore, the dew point can be estimated before even taking a fluid sample from the reservoir. Also, since  $C_2H_6$  as a predictor variable is absent from the model, the output of earlier versions of down-hole analyzer tools (which do not estimate the amount of  $C_2H_6$ ) may be used to predict the dew point pressure.

It is envisioned that the calculated dew point from this model could be used in several applications. First, it can be employed as a form of quality control to ensure the sampling procedure takes fluid samples with down-hole pressure above the dew point pressure (for more accurate sampling). This can serve as a confirmation to the current operational procedure of establishing dew point (by pumping out until liquid appears and is sensed by the tool sensors). This is particularly useful in low condensate gas ratio fluids. Secondly, it provides a quick estimate of dew point pressure that can help in any further estimation of phase behavior of gas wells for reservoir and production engineering applications. Thirdly, in cases having extensive down-hole data for multi-wells in the same reservoir, the calculated dew point pressure can be used to quality control the down-hole data. In this application, the trend of the calculated dew point pressure will be checked to see whether it follows the expected increase with depth trend. This estimation can also be used in confirming reservoir compartmentalization.

## Conclusions and recommendations

The objective of this paper was to introduce a dew point pressure correlation based on down-hole fluid analysis data. We used outputs of an existing tool in the industry to guide us through the selection of the inputs for the model. Put simply, we are introducing a quick model that the industry can use for the estimation of dew point pressure based on simple measured data. Our study proceeded as follows.

- The extensive literature review to identify all dew point pressure estimation models, classifying them into two groups, comparing the performance of each, and then suggesting which one performs better based on our extensive database.
- Building a model based on a small group of information-dependent variables that are measured from down-hole fluid analyzer tools, after deleting non-effective variables such as  $C_2H_6$  from the pool of independent variables.
- Testing and validating the model based on randomly selected data sets from our database.

A single best linear regression model that includes pairwise interactions was arrived at for the well data, by using a sophisticated statistical model selection criterion (AICc). We think the proposed model arrived at in this analysis is the best of its kind in the industry nowadays. A comparison of our proposed model versus published ones (although published models are based on surface data while our new model is based on down-hole data) shows similar results in terms of accuracy in predicting dew point pressure values. As a final recommendation, more refined models could be proposed in future work taking into account the collection of more data.

**Acknowledgements** The authors would like to acknowledge kind assistance through fruitful discussion with Rami Ahmed of Schlumberger, and Shabbir Hussain Shah Syed of Shell (STOS). We thank two anonymous referees for their constructive suggestions. The authors thank Hani Elshahawi, GameChanger; Formation Testing and Sampling Principal Technical Expert from Shell.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

- Al-Dhamen M, Al-Marhoun M (2011) New correlations for dew-point pressure for gas condensate. Presented at the SPE Saudi Arabia section Young Professionals Technical Symposium held in Dhahran, Saudi Arabia. SPE-155410
- Betancourt SS, Fujisawa G, Mullins OC (2004) Exploration applications of downhole measurement of crude oil composition and fluorescence. Presented at the Asia Pacific Conference on Integrated Modelling Asset Management held in Kuala Lumpur, Malaysia. SPE-87011
- Betancourt SS, Dubost FX, Mullins OC, Cribbs ME, Creek JL, Mathews SG (2007) Predicting downhole fluid analysis logs to investigate reservoir connectivity. Presented at the international petroleum Technology Conference held in Dubai, U.A.E. SPE-11488
- Burnham KP, Anderson DR (2002) Model selection and multi-model inference: a practical information-theoretic approach, 2nd edn. Springer, New York
- Elshahawi H, Hows M, Dong C et al (2007) Integration of geochemical, mud-gas, and downhole-fluid analysis for the assessment of compositional grading-case studies. Presented at the 2007 SPE Annual technical Conference and Exhibition, Anaheim, California. SPE-109684
- Elsharkawy A (2001) Characterization of the plus fraction and predicting the dew-point pressure for gas condensate reservoirs. Presented at the 2001 SPE Western Regional Meeting held in Bakersfield, California, USA. SPE-68776
- Elsharkawy A (2011) Predicting the dew-point pressure for gas condensate reservoirs: empirical models and equations of state. Fluid Phase Equilibria 4900, Elsevier Science B.V
- Godwin ON (2012) A new Analytical Method for Predicting Dew-point Pressure for Gas Condensate Reservoirs. Presented at the 2012 SPE Nigerian Annual International Conference and Exhibition, Nigeria. SPE-162985
- Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York
- Humoud AA, Al-Marhoun MA (2001) A new correlation for gas-condensate dew-point pressure prediction. Presented at the 2001 SPE Middle East Oil Show held in Bahrain, SPE-68230
- Marruffo L, Maita J, Him J, Rojas G (2002) Correlations to determine retrograde dew-point pressure and C7+ percentage of gas condensate reservoirs on basis of production test data of eastern Venezuelan fields. Presented in Gas Technology Symposium held in Calgary, Alberta, Canada. SPE-75686
- McCain WD (1994) Heavy components Control Reservoir Fluid Behavior. Paper SPE 28214, Technology Today series JPT 746-750
- Mullins O, Elshahawi H, Matthew F, O'Keefe M, Vanuffellen S (2009) The impact of reservoir fluid composition variation and valid sample acquisition on flow assurance evaluation. Presented at the 2009 Offshore Technology conference held in Houston, TX. USA. OTC-20204
- Nemeth K (1966) A correlation of dew-point pressure with reservoir fluid composition and temperature. PhD Dissertation Texas A&M University, College Station, TX, USA
- Nemeth LK, Kennedy HT (1966) A correlation of dew-point pressure with fluid composition and temperature. Presented at SPE Annual Fall meeting held in Dallas, TX. USA. SPE-1477
- Olds RH, Sage BH, Lacey WN (1944) Volumetric and phase behavior of oil and gas from paloma field. Presented at Los Angeles meeting, Los Angeles, USA. SPE-945077
- Ovalle A, Lenn C, McCain W Jr (2005) Tools to manage gas condensate reservoirs: novel fluid property correlations based on commonly available field data. Presented at IPTC held in Doha, Qatar. IPTC 10320-PP
- R Core Team (2016) R Foundation for Statistical Computing. Vienna, Austria. <http://www.R-project.org>

- Shokir E (2008) Dew-point Pressure Model for Gas Condensate Reservoirs based on Genetic Programming. King Saud University. Presented at CIPC gas technology Symposium 2008 held in Calgary Alberta., SPE-114454
- Xian C, Dawoud A, Carnegie A et al (2006) Identification and characterization of transition zones in tight carbonates by downhole fluid analysis. Presented at 2006 Abu Dhabi International Petroleum Exhibition and conference held in Abu Dhabi, U.A.E., SPE-101257